CS 6603: AI, Ethics, and Society
Final Project

Angelo Scaria, Bikash Jha, Darrian Parker, Oluwatobi Akerele
ascaria8@gatech.edu, bjha32@gatech.edu, dparker74@gatech.edu,
oakerele3@gatech.edu

1 DATASET SELECTION

Selected Dataset: Kaggle - Students Performance Dataset - Source Link

Regulated Domain(s) pertaining to Dataset:

- Education

# of Dataset Observations: 2,392

# of Dataset Variables: 15

Dependent/Outcome Variables: 2

- GPA
- GradeClass

Variables associated with a Legally Recognized Protected Class: 3

- Age
- Gender
- Ethnicity

Legal Precedence pertaining to each Protected Class:

- Age (Age Discrimination in Employment Act of 1967)
- Gender (Equal Pay Act of 1963; Civil Right Acts of 1964, 1991)
- Ethnicity (Civil Right Acts of 1964, 1991)

2 DATASET EXPLORATION

2.1 Protected Class Variables and Associated Subgroups

The following table outlines the subgroups associated with the protected class variables gender and ethnicity in the dataset.

Table 1 - Protected Class Variables and Associated Subgroups

Protected Class Variable	Subgroup
Gender	Male
	Female
Ethnicity	Caucasian
	African American
	Asian
	Other

2.2 Protected Class Variables and Associated Subgroups Discretization

The following table illustrates discretization of the subgroups associated with the protected class variables identified in 2.1.

Table 2 - Discretization of Associated Subgroups into Numerical Values

Protected Class Variable	Subgroup	Numerical Value
Gender	Female	0
	Male	1
Ethncity	Caucasian	0
	African American	1
	Asian	2
	Other	3

2.3 Selected Protected Class Variables for Analysis

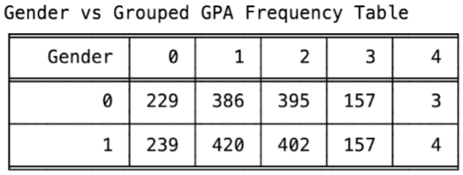
Following are the selected protected class:-

- Gender
- Ethnicity

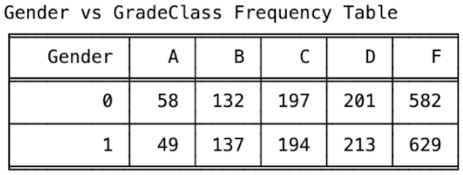
2.4 Frequency of Protected Class Subgroups within Dataset

The following table details the frequency of occurrence between each identified subgroup within the protected class. The mapping is as follows
GradeClass is mapped as follows
0=A,1=B,2=C,3=D,4=F
GPA is mapped as follows
0-1 gpa=0
1-2 gpa=1
2-3 gpa=2
3-4 gpa=3
4 gpa=4

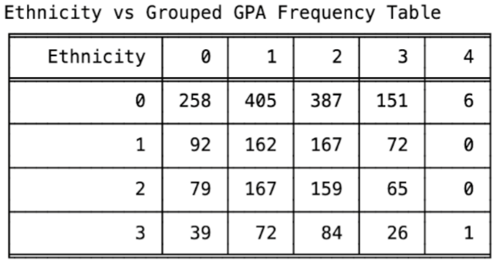
Table 3 – Protected Class Variables vs Outcome Variables (Frequency Tables)


Gender vs Grouped GPA Frequency Table

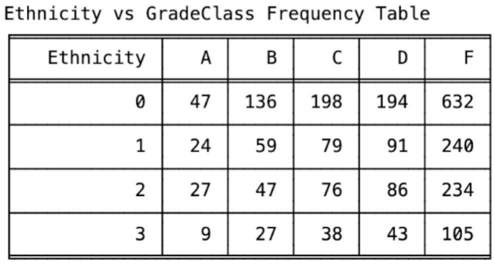
Gender	0	1	2	3	4
0	229	386	395	157	3
1	239	420	402	157	4


Gender vs GradeClass Frequency Table

Gender	A	B	C	D	F
0	58	132	197	201	582
1	49	137	194	213	629


Ethnicity vs Grouped GPA Frequency Table

Ethnicity	0	1	2	3	4
0	258	405	387	151	6
1	92	162	167	72	0
2	79	167	159	65	0
3	39	72	84	26	1


Ethnicity vs GradeClass Frequency Table

Ethnicity	A	B	C	D	F
0	47	136	198	194	632
1	24	59	79	91	240
2	27	47	76	86	234
3	9	27	38	43	105

2.5 Graphing Frequency of Protected Class Subgroups

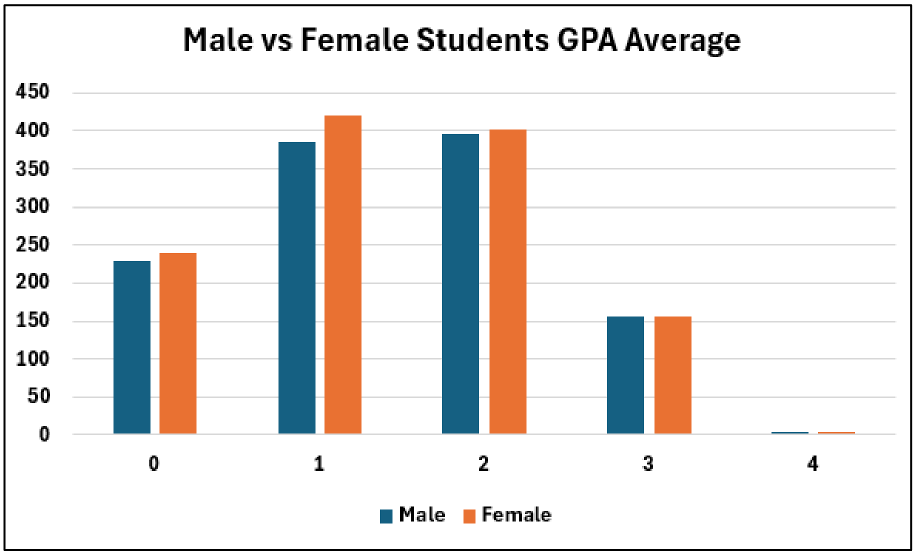


Figure 1 – Male vs Female Students GPA Average

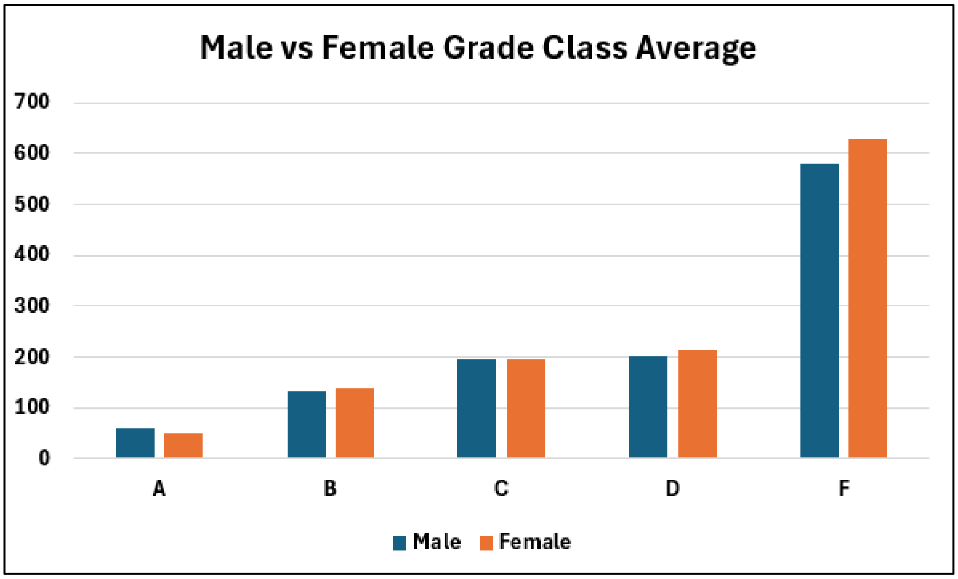


Figure 2 – Male vs Female Students Grade Class Average

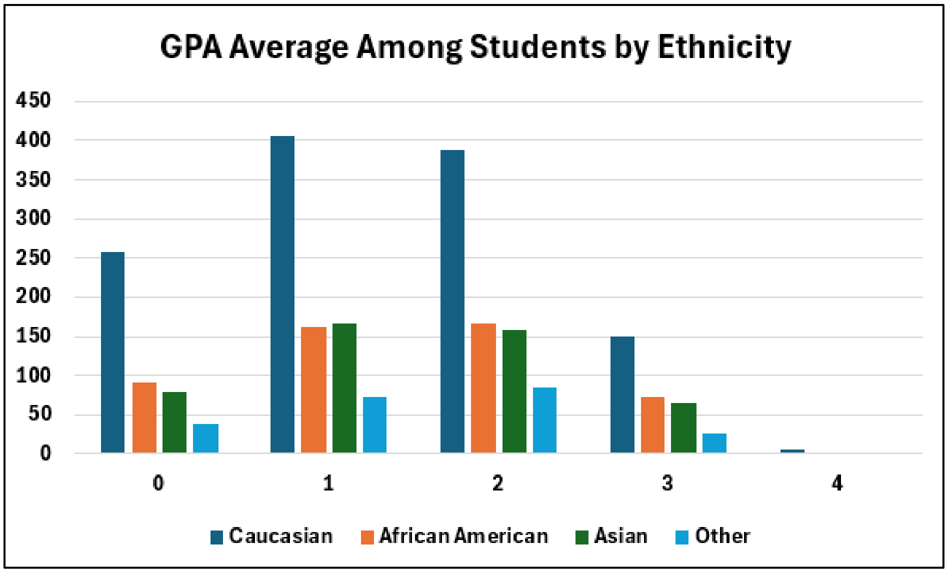


Figure 3 – GPA Average Among Students by Ethnicity

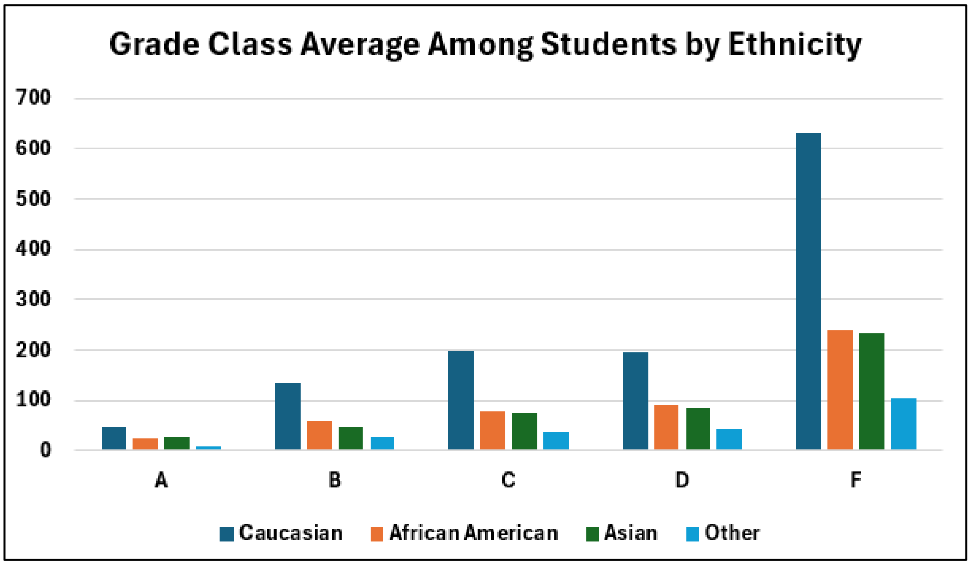


Figure 4 – Grade Class Average Among Students by Ethnicity

3 FAIRNESS METRICS, MITIGATING BIAS ALGORITHM

3.1 Privileged and Unprivileged groups for each Protected Class

Privileged group for gender:
Male[1]

Unprivileged group for gender:
Female[0]

Privileged group for ethnicity:
Caucasian[0]

Unprivileged group for ethnicity:
African American[1], Asian[2], Other[3]

3.2 Fairness metric algorithms for each protected class in the dataset

In this project we will be using following fairness metric algorithms:

- Disparate Impact
- Statistical Parity Difference

Below is the table with fairness metric algorithms for each protected class

Table 4 – DI and SPD Metrics for Each Protected Class by Outcome Variable

Protected Class	Fairness metrics algorithm	Value(GradeClass)	Value(GPA)
Gender	Disparate Impact	1.06691	1.03796
	Statistical Parity Difference	0.0101837	0.0050
Ethnicity	Disparate Impact	1.07422	1.06398
	Statistical Parity Difference	0.0112536	0.00832

3.3 Pre-processing bias mitigation algorithm

Reweighting was used as the pre-processing bias mitigation algorithm to transform the original dataset

3.4 Computing fairness metrics on the transformed dataset

Below are the results after computing fairness metrics on the transformed dataset.

Pre-processing bias mitigation algorithm used: Reweighting from AIF360.

Table 5 – DI and SPD Metrics for Each Protected Class Post-Bias Mitigation

Protected Class	**Fairness metrics algorithm**	**Value(GradeClass)**	**Value(GPA)**
Gender	Disparate Impact	1	1
	Statistical Parity Difference	0.0000000000000000277556	0
Ethnicity	Disparate Impact	1	1
	Statistical Parity Difference	0.0000000000000000277556	-0.0000000000000000277556

4 MITIGATING BIAS TECHNIQUE

We chose to use an 80-20 split to the dataset where 80% of the data lies in the training set and 20% of the dataset lies in the testing set. Our execution can be found in our .ipynb file. Furthermore, we chose Random Forest to be our classifier in this step.

Below, in table format, are the differences in the outcomes for privileged versus unprivileged group for each fairness metric calculated, per protected class, after training the classifier on the original and transformed datasets.

Table 6 – Outcomes for DI Metric on Gender Protected Class

	Disparate Impact Protected Class - Gender	**Change compared to previous**
Original Dataset	1.06691	NA
After Transforming Dataset	1	Minimal positive change
After Training Classifier on Original Dataset	1.19724	Negative change
After Training Classifier on Transformed Dataset	1.19724	No change

Table 7 – Outcomes for SPD Metric on Gender Protected Class

	Statistical Parity Protected Class - Gender	**Change compared to previous**
Original Dataset	0.0101837	NA
After Transforming Dataset	2.77556E-17	Minimal positive change
After Training Classifier on Original Dataset	0.0287180	Minimal negative change
After Training Classifier on Transformed Dataset	0.0287180	No change

Table 8 – Outcomes for DI Metric on Ethnicity Protected Class

	Disparate Impact Protected Class - Ethnicity	**Change compared to previous**
Original Dataset	1.07422	NA
After Transforming Dataset	1	Minimal positive change
After Training Classifier on Original Dataset	1.249004	Negative change
After Training Classifier on Transformed Dataset	1.074224	Positive change

Table 9 – Outcomes for SPD Metric on Ethnicity Protected Class

	Statistical Parity Protected Class - Ethnicity	**Change compared to previous**
Original Dataset	0.0112536	NA
After Transforming Dataset	2.77556E-17	Minimal positive change
After Training Classifier on Original Dataset	0.0349479	Minimal negative change
After Training Classifier on Transformed Dataset	0.0112536	Minimal positive change

5 ANALYSIS

5.1 Members of Project Team

- Angelo Scaria
- Bikash Kumar Jha
- Darrian Parker
- Oluwatobi Akerele

5.2 Graph Fairness Metric Results

The following charts detail the Disparate Impact and Statistical Parity Difference metrics from the original dataset, transformed dataset, and the dataset trained with a classifier.

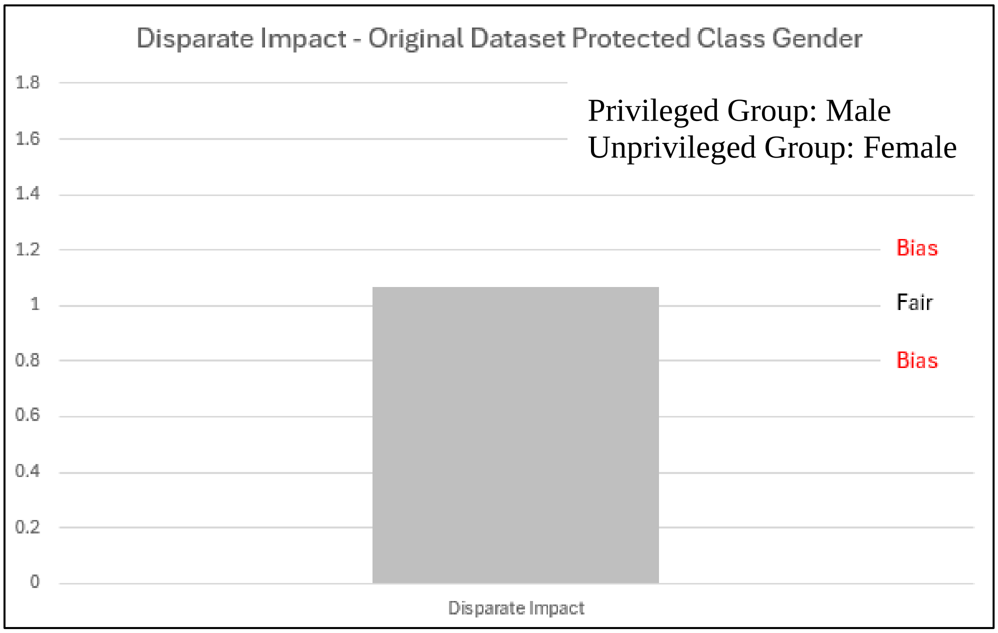


Figure 5 – DI Metric for Protect Class Gender

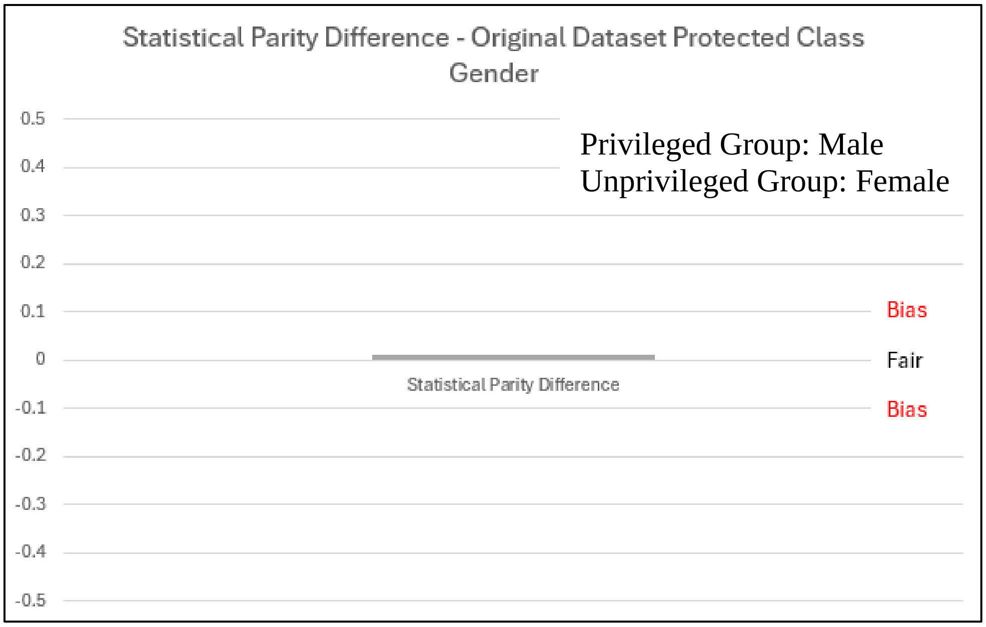


Figure 6 – SPD Metric for Protected Class Gender

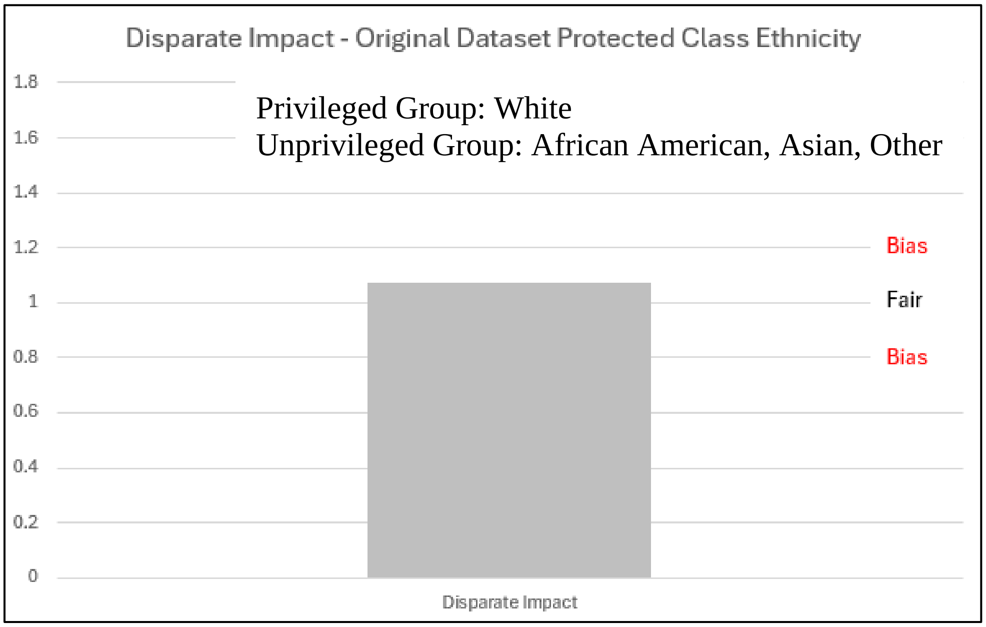


Figure 7 – DI Metric for Protected Class Ethnicity

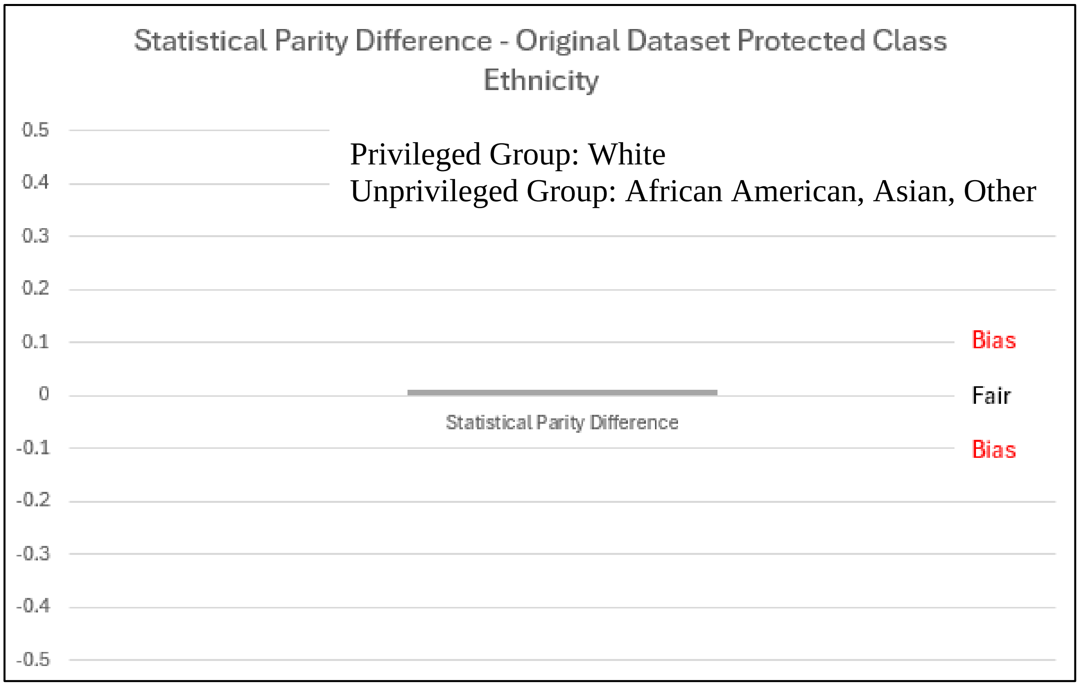


Figure 8 – SPD Metric for Protected Class Ethnicity

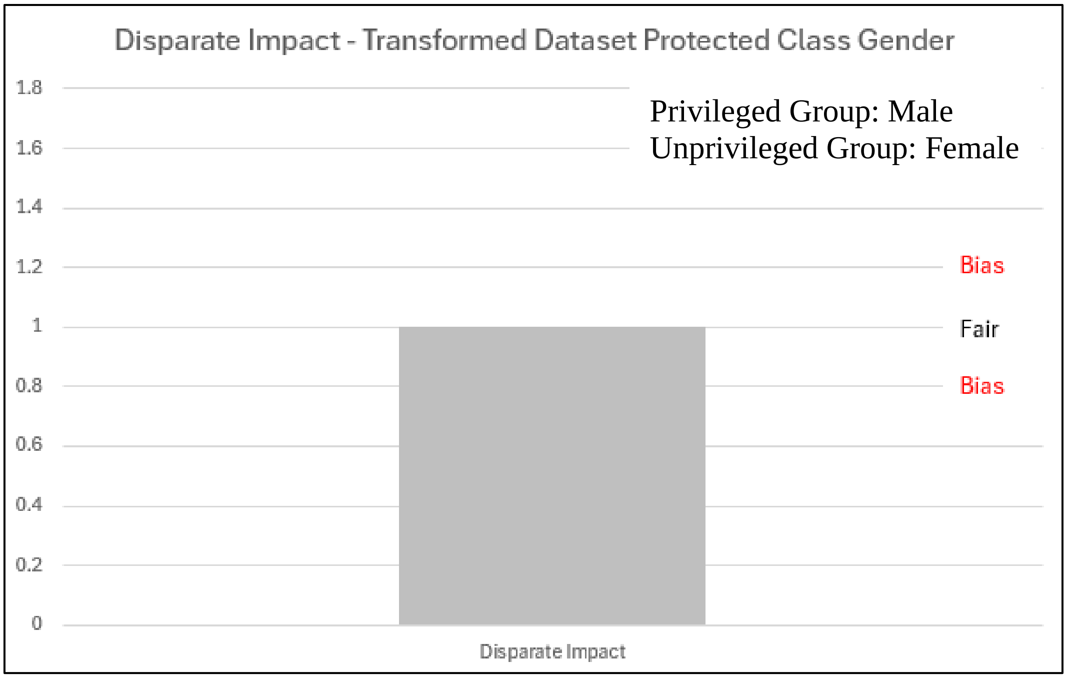


Figure 9 – DI Metric for Protected Class Gender (Transformed Dataset)

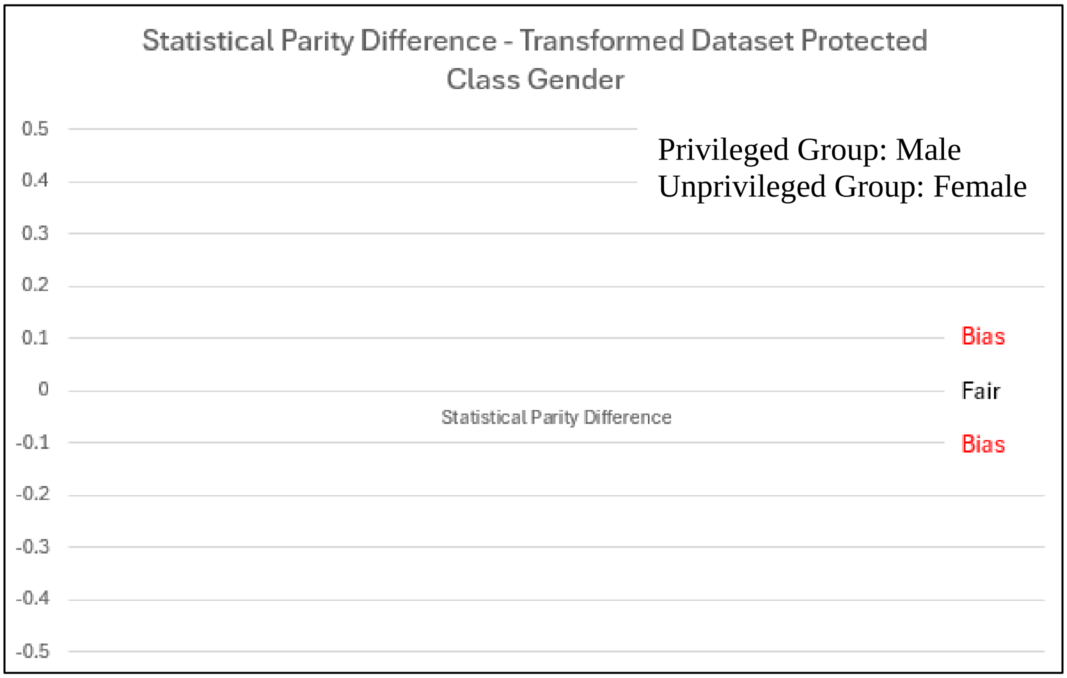


Figure 10 – SPD Metric for Protected Class Gender (Transformed Dataset)

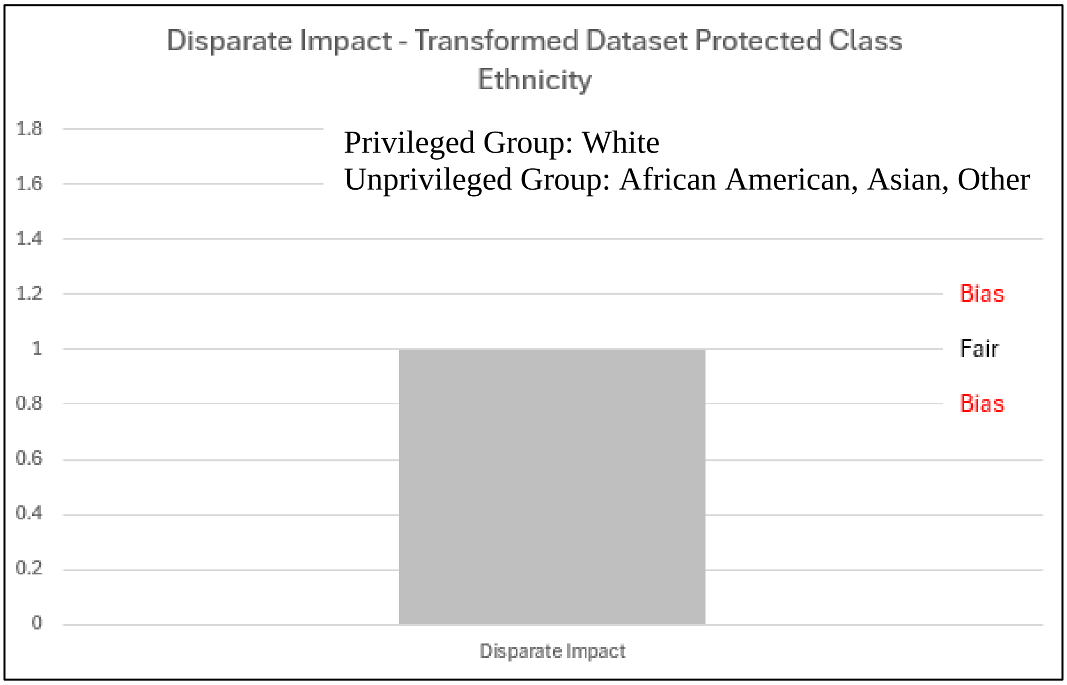


Figure 11 – DI Metric for Protected Class Ethnicity (Transformed Dataset)

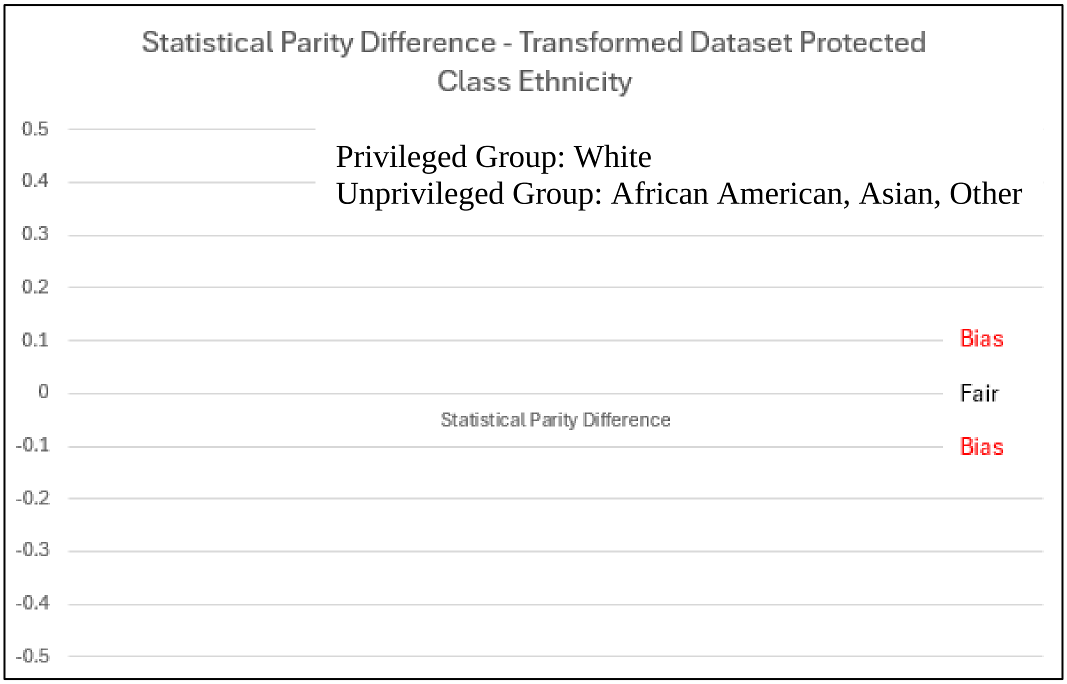


Figure 12 – SPD Metric for Protected Class Ethnicity (Transformed Dataset)

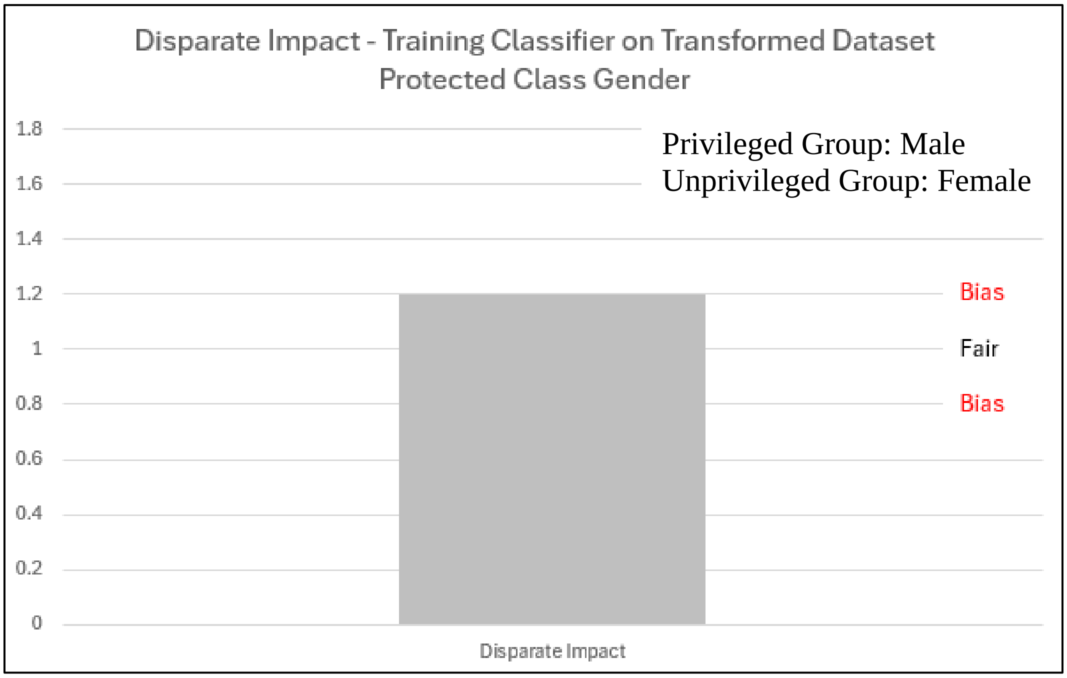


Figure 13 – DI Metric for Protected Class Gender (Dataset with Training Classifier)

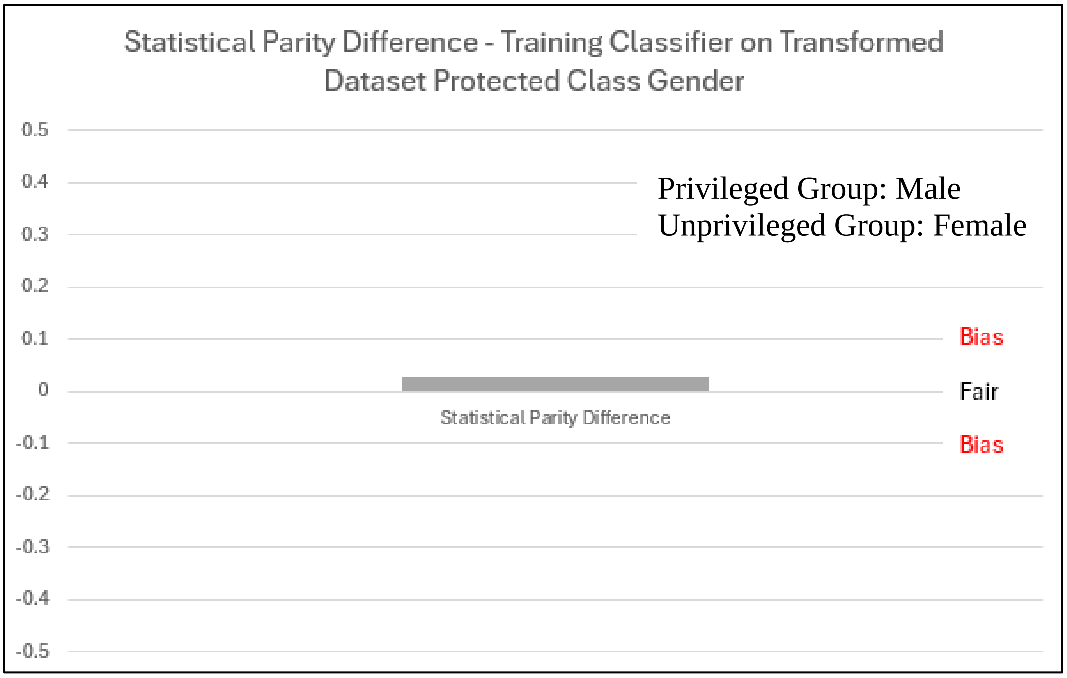


Figure 14 – SPD Metric for Protected Class Gender (Dataset with Training Classifier)

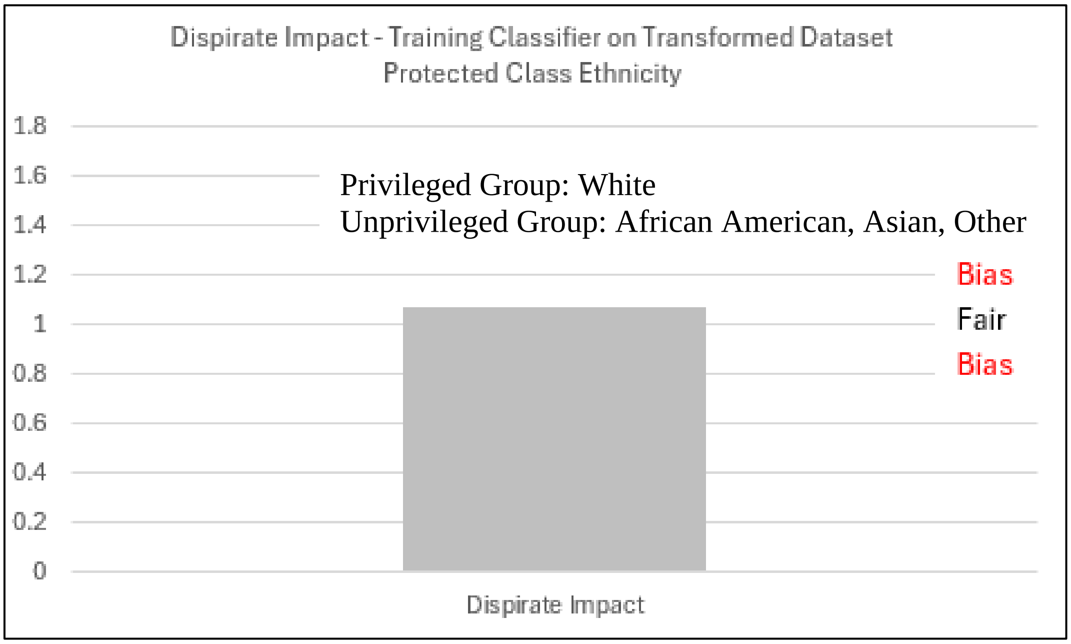


Figure 15 – DI Metric for Protected Class Ethnicity (Dataset with Training Classifier)

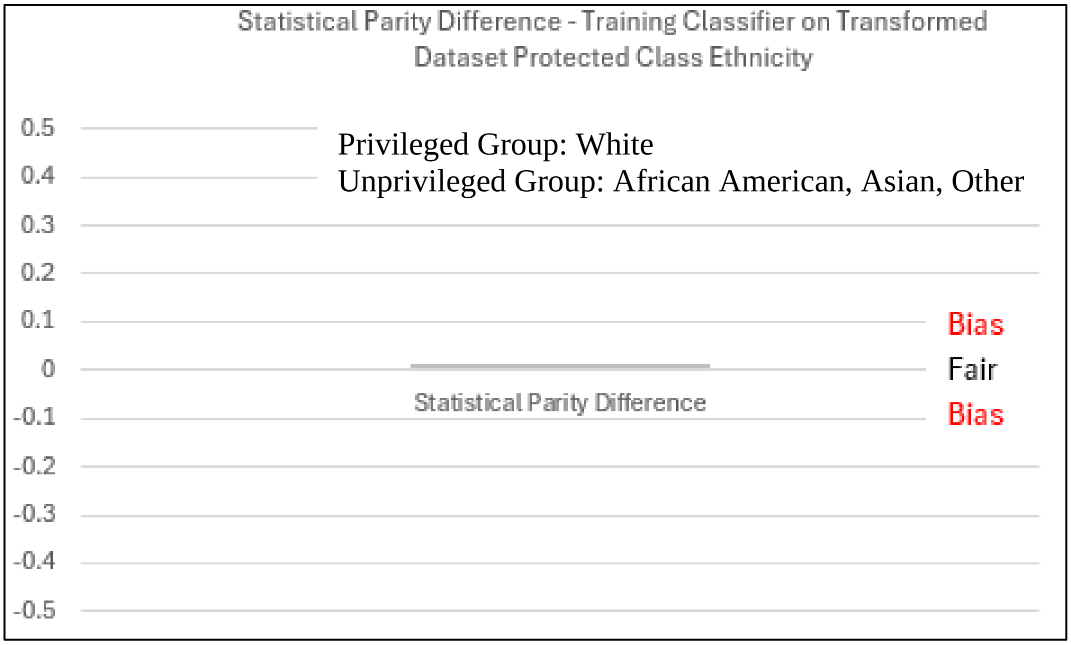


Figure 16 – SPD Metric for Protected Class Ethnicity (Dataset with Training Classifier)

5.3 Best Fairness Metric Analysis

Of the two metrics selected, Disparate Impact (DI) and Statistical Parity Difference (SPD), Disparate Impact is the best metric considering this analysis.

As SPD has a smaller delta between the original dataset metrics and the metrics after Reweighing, it is more difficult to draw definitive conclusions. For both Gender and Ethnicity Protected Class analysis there is marginal change among the original, transformed, and training classified datasets. Using SPD for Gender with outcome variable GradeClass as an example at 0.0101, 2.77556E-17, and 0.0287 respectively. This tells us our data is fair by industry standards, although not absolute.

Using the same example, DI ratio for Gender with outcome variable GradeClass, throughout the analysis the ratio changes from 1.07, 1, and 1.20 for the original, transformed, and training classified datasets respectively. This shows a clearer advantage towards the unprivileged group Female, that is mitigated to an ideal value of 1 using the Reweighing method. This metric further shows the Female subgroup having an advantage upon using a training classifier.

For these reasons, the Disparate Impact ratio is best used when determining bias and fairness among the Protected Class variables.

5.4 Individual Team Member Analysis

Angelo Scaria

- Using the Reweighing method did mitigate bias as needed to an almost perfect amount. The initial disparate impact value with respect to the gender protected class and GradeClass outcome variable was 1.07 and changed to 1.0 after reweighing. The initial statistical parity difference value with respect to the gender protected class and GradeClass outcome variable was 0.01 and changed to 2.8e-17 after reweighing. The initial disparate impact value with respect to the ethnicity protected class and GradeClass outcome variable was 1.074 and changed to 1.0 after reweighing. The initial statistical parity difference value with respect to the ethnicity protected class and GradeClass outcome variable was 0.011 and changed to 2.8e-17 after reweighing. Results were comparable for the GPA outcome variable too.
- The groups that received the most positive advantage from reweighing are certainly the unprivileged groups in the race class (African American, Asian, Other) because before reweighing, the disparate impact value was 2.24 and it changed to 1.0 after the transformation. Also, another group that received a positive advantage was the privileged group of the gender class (male) as the disparate impact value in respect to GPA was 0.81 before reweighing and changed to 1.0 after reweighing.
- A group that received a negative advantage from reweighing is the privileged group in the race class (White) because before reweighing, the disparate impact value was 2.24 and it changed to 1.0 after the transformation. Also, another group that received a negative advantage was the unprivileged group of the gender class (female) as the disparate impact value in respect to GPA was 0.81 before reweighing and changed to 1.0 after reweighing.
- The issues that would arise if I used these methods to mitigate bias include the groups that were disadvantaged treatment being unhappy with these results. While these methods will result in a fairer model, if another dataset were used, performance could have been negatively impacted.

Bikash Kumar Jha

- The reweighting method substantially reduced bias and improved fairness in the data. This was evidenced by a considerable decrease in the Disparate Impact (DI) and Statistical Parity Difference (SPD) values following modification. The DI values for both gender and ethnicity protected classes were adjusted to one, showing that the reweighting procedure treated privileged and unprivileged groups equally.

Furthermore, the SPD values approached zero, indicating that the disparities in positive outcomes between groups were reduced. This consistent improvement across several criteria shows that the reweighting strategy improved the model's fairness.

- Yes, certain groups benefited from the reweighting method. Among the ethnically protected classes, the unprivileged groups, which included African American, Asian, and Other ethnicity subgroups, profited greatly because their initial DI values were decreased to one, signifying equitable treatment following reweighting. Similarly, the privileged group in the gender class, particularly males, witnessed an improvement in GPA. Males had an initial DI value of 0.81, indicating a disadvantage, but after reweighting, this value changed to 1, indicating that the reweighting procedure also leveled the playing field for this group.
- Yes, the reweighting method disadvantaged certain populations. The Caucasian subgroup, which was the privileged group within the ethnicity class, had its DI value reduced from 2.24 to 1. This suggests that their early advantage was neutralized by the reweighting procedure. Similarly, within the gender class, females' DI value related GPA decreased from 0.81 to 1, showing that they had lost their pre-reweighting modest edge. These changes, while enhancing fairness, resulted in formerly favored groups losing their preferential status.
- Using these bias mitigation strategies could result in a number of concerns. For starters, there is the possibility of shifting bias, in which the strategy equalizes treatment between groups while disadvantageously affecting previously advantaged groups, producing unhappiness among them. Second, there is frequently a trade-off between justice and model performance; maintaining fairness may occasionally lower forecast accuracy, which is crucial in real-world applications. Third, ethical problems come into play, especially when historically disadvantaged groups, which may naturally do better, lose their advantage due to bias mitigation. Fourth, stakeholder satisfaction is a worry because different groups may have varied perspectives of what constitutes justice, potentially leading to reaction from those who believe the changes have been unfairly applied. Finally, bias mitigation methods are context sensitive, which means that a methodology that works well in one dataset or domain may not perform as well in another, necessitating adapted approaches for diverse contexts.

Darrian Parker

- The Reweighing approach was successfully able to both mitigate bias and increase fairness in the selected metrics, Disparate Impact and Statistical

Parity Difference. While both metrics were already deemed fair by industry standards (falling within 0.8 and 1.25 for Disparate Impact; falling within -0.1 and 0.1 for Statistical Parity Difference), these metrics still saw marginal improvement.

- As measured by the original Disparate Impact ratio for Gender, 1.067, the Female subgroup did have a slight advantage over the Male subgroup as the ratio was greater than 1. This was mitigated using the Reweighing bias mitigation technique and bringing this ratio to an absolute value of 1.
- In reviewing the metrics from an ethical standpoint, one could argue that the Female subgroup disadvantaged gender. Their subgroup is the historically disadvantaged subgroup among genders, yet they had the advantage in this analysis of school performance. In this case or a real-world scenario, one might refrain from leveraging a bias mitigation technique if a historically disadvantaged subgroup benefits.
- The issue that arises when using bias mitigation techniques, is that one group will benefit at the expense of another. Removing bias may be viewed as an ethical approach, unless you are the subgroup that is giving up their advantage. This also brings on a question of whether bias mitigation is needed, rather than taking the raw results of the dataset to be truly representative of the findings.

Oluwatobi Akerele:

- An approach that seemed to mitigate bias in our analyses was the use of the Reweighing pre-processing bias mitigation algorithm and the transformation of the original dataset as a function of the GradeClass dependent variable. The bias mitigation approach led to a positive change in the Disparate Impact fairness metric for both protected class attributes, indicating that the model treated both privileged and unprivileged groups fairly and equally.
- After transforming the dataset, there was a reduction in the statistical parity difference from 0.0101837 to 2.77556E-17 for the Gender protected class, and from 0.0112536 to 2.77556E-17 for the Ethnicity protected class. This reduction indicates that the privileged group received a positive advantage from the bias mitigation.
- However, after training the classifier on the transformed dataset, there was an increase in statistical parity difference, with values up to 0.0287180 for the Gender protected class and values up to 0.0349479 for the Ethnicity protected class. This indicates a disadvantage for the privileged group, because it shows that the unprivileged group now has a higher favorable outcome rate compared to the privileged group.

- An issue that would arise from the approach to train the classifier on the transformed data, would be the perpetuation of existing disparities between the privileged and unprivileged groups. This is because we witness an increase in statistical parity difference, indicating that the unprivileged group gained higher favorable outcome rates after the bias mitigation approach.